Support to Building the Inter-American Biodiversity Information Network

Trust Fund #TF-030388

Review of experience in developing interoperable systems for international data management and sharing

(Document 10 (b))

July 2004

Support to Building IABIN (Inter-American Biodiversity Information Network) Project

Review of experience in developing interoperable systems for international data management and sharing

Project Background

The World Bank has financed this work under a trust fund from the Government of Japan. The objective is to assist the World Bank in the completion of project preparation for the proposed project Building IABIN (Inter-American Biodiversity Information Network) and for assistance in supervision of the project, once and if it is approved. The work undertaken covers three areas: background studies on key aspects of biodiversity informatics; direct assistance to the World Bank in project preparation; and assistance to the World Bank in project supervision. The current document is one of the background studies.

The work has been carried out by Nippon Koei UK, in association with the UNEP World Conservation Monitoring Centre.

Table of Contents

Annexes

Table of Tables

Table of Figures

Report Summary

This report reviews experience of developing interoperable systems for data management and sharing in biodiversity at the international level. At the time of reporting these types of interoperability initiatives are in flux as a new generation of Web based technologies start to become available. In particular the service-based architecture of Web Services, which provide self-describing modular applications, has come to the fore. The underlying technologies and standards of Web Services are reviewed together with biodiversity related standards for information encoding and transfer. Lastly, a selection of biodiversity information sharing initiatives, which are based on both traditional and more recent Web Services architectures, are reviewed.

CHAPTER 1 INTRODUCTION

This report considers and reviews experience at the international level in developing interoperable systems for data management and sharing. Report 7.2 is concerned with using metadata (“data about data”) to aid information resource discovery (finding data/information), whilst this report is concerned with managing and sharing data/information. These activities are clearly related, but the division between them can be unclear, mainly because one person’s metadata can be another person’s data. Report 7.2 is concerned primarily with “discovery metadata”, finding things at a high and fairly generic level. This report is also concerned, in part at least, with metadata, but of a different granularity.

There have been many initiatives and standards for exchanging and sharing biodiversity data. In the same way that searching for books in the library is a generic task, data management and sharing is also a generic task, although different subject domains, even within the biodiversity arena, will have important subject specific requirements. Past initiatives have, naturally, used the tools of the day. However, with the emergence of Web Services as a generic standards-based framework, there is now the opportunity to build interoperable systems more easily, and there is already a clear trend to do this.

Biodiversity information covers a broad church relating to taxonomy, species, ecosystems, protected areas and responses. Information management and sharing requirements, in each of these areas, necessarily differ.

CHAPTER 2 KEY PRINCIPLES OF DEVELOPING INTEROPERABLE DATABASES

2.1 Informatics Standards – Why Use Them?

Adherence to information standards aids reuse, portability and interoperation between collaborators. The WWW works by all browsers adopting HTTP, a simple example that illustrates the power of using a standard for presenting information. Standards are important, indeed essential, in developing interoperable systems for content transfer, analysis and presentation. This report deals with two types: generic standards for information sharing, and standards that are specific to biodiversity informatics.

2.2 Database Design and Data Models

Database design and data modelling is a well-developed area of informatics and uses formal methods to ensure efficient storage, query and retrieval, and data integrity. The ubiquitous relational model has been adopted by most mainstream database providers, and standardisation on the Structured Query Language (SQL) means that a degree of interoperability already exists between systems. However, most databases have, been designed for a specific purpose and will, in all likelihood, not be directly comparable or compatible with similar systems developed in other institutions, even if they fulfil a similar role.

For many years entity-relationship (ER) modelling was the primary tool in database design. Entity-relationship modelling breaks a data management task into entities – an object about which we wish to store data, the entity’s attributes (data), and the relationships between the entities. Entity-relationship modelling has given way to design using the Unified Modelling Language (UML), which performs a similar design task using object orientated techniques and a different notation.

Space does not permit further discussion of database design, but it should be stressed that in the realm of biodiversity informatics, database design and management are crucial issues. Biodiversity data, both historical and current, are very valuable resources, and managing them should be seen as an investment not a cost. Too frequently scientists, whose backgrounds are not in informatics, use informatics tools in inappropriate ways. For example, spreadsheet programs are often used to store data that are more appropriately stored in a database. Spreadsheets store data in matrices that do not reflect the structure of the data. Even when data are stored in a database the system is frequently badly designed, constraining the utility and accessibility of data, and even resulting in information loss. Data collection exercises should commence with a data-modelling phase, followed by design and database implementation.

2.3 Data Management

The data management cycle consists of a number of distinct phases, identified by BCIS (2000) as:

- Data pre-collection;
- Data collection;
- Data processing;
- Data submission;
- Data management.

The cycle may be short, with data being collected, analysed and discarded, or may be long with data, and specimens, being collected, analysed and curated for many, perhaps hundreds of, years. It costs time and money to collect data and they may have a high intrinsic value. Indeed some organisations' business models are based on their data holding, so data management becomes a central issue.

Data should be described and catalogued using metadata and should not be considered complete without it.

Appropriate tools should be used for data collection, storage and management. Despite their wide use, spreadsheet programs are not a suitable framework in which to store data, as they do not have the required structures to assure efficient data storage and maintenance of integrity.

Computer based systems are now ubiquitous and have many advantages. However, computer systems change rapidly and much valuable computer data has been lost through degradation of magnetic media and changing media formats.

A full discussion of data management procedures is beyond the scope of this report and the reader is referred to Volume 2 of the BCIS HowlkjdfFramework for Information Sharing, BCIS (2000). This report is available online at http://www.biodiversity.org

2.4 Data Policy

Information sharing initiatives are generally based on the assertion that ownership and custodianship of data and information is important and should be clearly defined and documented. The distributed data model assists in defining ownership boundaries by enabling data providers to expose only that information which they wish to share, and providing access constraints where necessary.

CHAPTER 3 STANDARDS AND PROTOCOLS FOR INTEROPERABILITY

3.1 Introduction

Interoperability is defined by the HyperDictionary (www.hyperdictionary.com) as:

"... the ability to exchange and use information (usually in a large heterogeneous network made up of several local area networks)"

Interoperability, based on standards, is the key to information sharing and one of the primary objectives of the IABIN network.

This chapter is split into three main parts and deals first with generic protocols for data exchange, followed by protocols specific to biological information, and lastly issues relating to the design of systems for data sharing. Each of the protocols discussed is reasonably self-contained, but many of them are not very useful in isolation. They do, however, fit together to make a powerful nexus for data interoperability and information sharing.

3.2 Standards Bodies

There are many Standards bodies and Tanenbaum's adage that "the nice thing about standards is that there are so many of them to choose from", is relevant in this context. The major Standards bodies are described briefly below. It should be remembered that behind each of these are the individuals and organisations that have come together with the desire to standardise their activities, with the goal of interoperability being a driving force.

3.2.1 The World Wide Web Consortium - W3C

The World Wide Web Consortium, or W3C, was founded in 1994 by Tim Berners-Lee, the inventor of the WWW, to develop a vision and common protocols for evolving an interoperable WWW. The Consortium's goals are to provide universal access across cultural and language barriers, to develop a software environment that maximises utility and to do so cognisant of the legal, commercial and social issues (Source W3C web site).

One of the major activities of the W3C is in publishing W3C Standards, or Recommendations, on web technologies, based on a consensus process. W3C has published more than 40 Recommendations, some of which are of particular relevance to biodiversity information sharing. The breadth and depth of these is formidable and it can be difficult to obtain a rational overview. However, the component building blocks of the Web do form a coherent picture, as do the various biodiversity standards, based upon W3C Recommendations, such as EML and ABCD. Interoperability is at the heart of the W3C and using its new generation of tools, interoperable systems for biodiversity information sharing can be constructed.

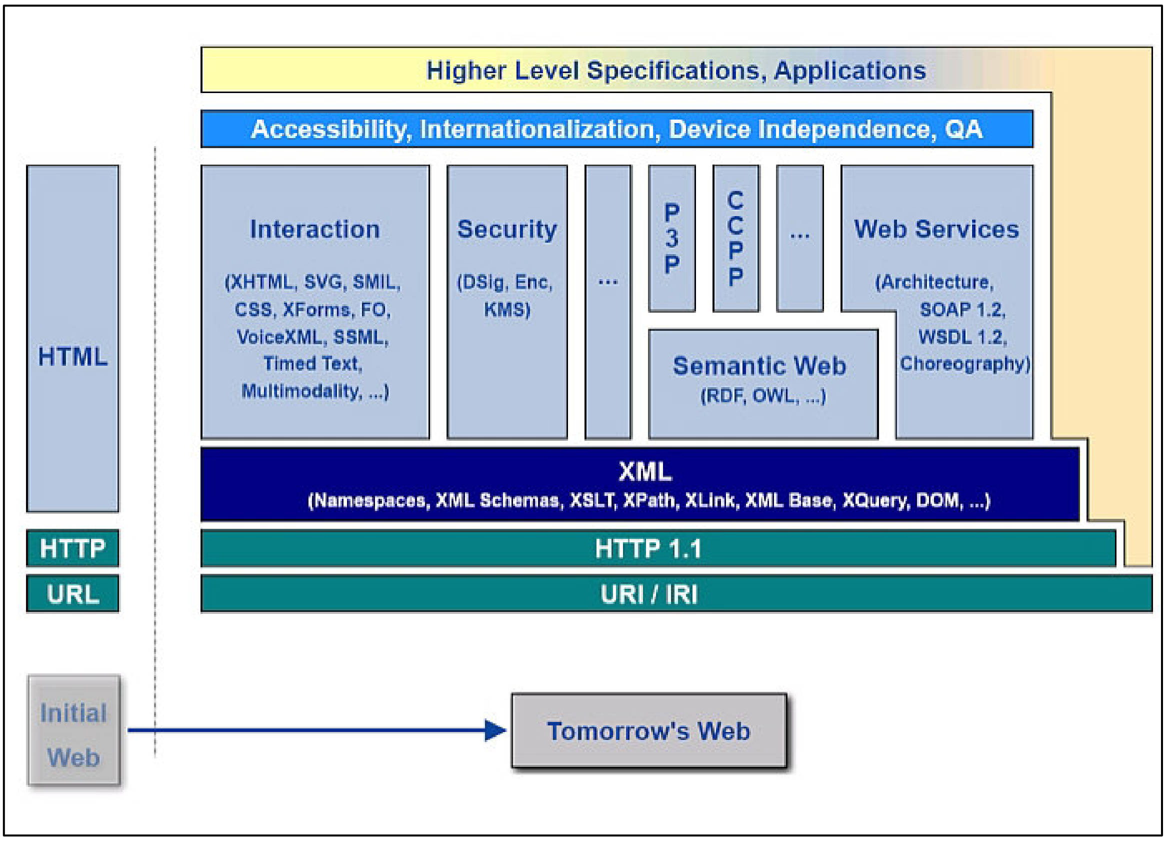

Higher Level Specifications, Applications
Accessibility, Internationalization, Device Independence, QA
HTML
Interaction
(XHTML, SVG, SMIL, CSS, XForms, FO, VoiceXML, SSML, Timed Text, Multimodality, ...)
Security
(DSig, Enc, KMS)
...
P3P
CCPP
...
Web Services
(Architecture, SOAP 1.2, WSDL 1.2, Choreography)
Semantic Web
(RDF, OWL, ...)
XML
(Namespaces, XML Schemas, XSLT, XPath, XLink, XML Base, XQuery, DOM, ...)
HTTP
HTTP 1.1
URL
URI / IRI
Initial Web
Tomorrow's Web

Figure 1. Evolution of WWW Technologies (Source http://w3c.org)

The left hand section of Figure 1 shows how the initial WWW was based on HTML, HTTP and URLs, and it was this that produced the revolution in information. Based on the new technologies of XML, XML Schema Definition, RDF, Web Services, etc. this revolution will both continue and accelerate. The component blocks in Figure 1 form a coherent whole. The biodiversity related data interoperability initiatives described below generally map onto one or other of these W3C components or lead towards them. XML has a cross cutting role in “Tomorrow’s Web” which will become evident below.

W3C Recommendations are available at no cost from http://w3c.org

3.2.2 International Organization for Standardisation – ISO

There are many national organisations concerned with standardisation at the national level. The International Organization for Standardization (ISO) is an NGO network of the national standards institutes of 148 countries. ISO in no way circumvents the work of the W3C. It does, however, publish a number of relevant standards, in particular those on geospatial metadata, as defined by ISO 19115 (a.k.a FGDC) and ISO 15836 (a.k.a. DCMI). The ISO 14000 Standards Family for Environmental Management are not explicitly concerned with biodiversity, but have an indirect bearing through general environmental management and best practice and their impact on biodiversity. ISO makes a charge for its standards documentation.

3.2.3 Open GIS Consortium (OGC)

The Open GIS Consortium (OGC) is a "non-profit international trade association" that defines interoperability standards for geoprocessing, based on open programming interfaces. Deliverables 2.7-2.9 deal specifically with Geographic Information Systems (GIS) so further discussion is not necessary here. However, it should be pointed out that GIS has a sound foundation in database systems and has been a driving force in recent years in standardisation and interoperability issues, particularly through the OGC and its members. Many GIS vendors now support OGC interoperability standards such as the Geographic Markup Language (GML) based on XML. The OGC is now developing in the framework of Web Services, described below, and has coined OGC Web Services (OWS) as "... a web of geoprocessing services that can be connected in dynamic, open interoperable chains to create dynamic applications".

3.2.4 Taxonomic Database Working Group - TDWG

The Taxonomic Database Working Group (TDWG) is a working group of the International Union of Biological Sciences and has been working since 1994 on defining standards for taxonomic databases, including botany, zoology and microbiology. The standards defined by TDWG and listed on its web site include:

- Authors of plant names;
- International transfer format for botanic garden plant records;
- Economic botany data collection standard;
- XDF – a language for the definition and exchange of biological data sets;
- Floristic regions of the world;
- World geographical scheme for recording plant distributions;
- Herbarium information standards and protocols for interchange of data;
- Plant names in botanical databases;
- Plant occurrence and status scheme (POSS).

The TDWG web site is not very current, and access to the standards is not always transparent. Furthermore, its activities, starting in 1994, have overlapped with the emergence of the W3C Recommendations, so it has presumably been difficult to align with these as they have emerged. However, the joint CODATA/TDWG task group on Access to Biological Collection Data (ABCD) is firmly in the mainstream of XML Schema Definition activity, although there is no link to it from the TDWG web site.

Further discussion on ABCD can be found below and on the ABCD site at:
http://bgbm3.bgbm.fu-berlin.de/TDWG/CODATA/default.htm
Further information on TDWG can be found at http://www.tdwg.org

3.3 Generic Standards and Protocols

3.3.1 JDBC & ODBC

Open DataBase Connectivity (ODBC) is a standard of the SQL Access Group that defines a standard Structured Query Language (SQL) interface to a wide range of database systems. ODBC translates ODBC compliant SQL into the specific SQL flavour used by each supported database system, thereby providing a standard interface to them all. These database systems may be personal desktop systems, open source or large commercial database systems.

Java DataBase Connectivity (JDBC) is part of Sun Microsystem’s Java Development Kit and provides access to generic SQL database systems through a standard programming interface, or API. It therefore fulfils a similar role to ODBC. A bridge between JDBC and ODBC is available.

Both JDBC and ODBC enable software programs to be written that access the contents of a very wide range of databases that may be either local applications, such as MS Access, or remote database servers on a local area network or the Internet. Both JDBC and ODBC provide interoperability, through a standard SQL, to heterogeneous database systems. They do not provide a framework for information sharing *per se*, but they do provide part of the glue to build interoperable frameworks.

3.3.2 eXtensible Markup Language - XML

The eXtensible Markup Language (XML) is a recommendation of the W3C (see above) that defines a simple meta-language to mark up data with human readable tags. In the same way that HTML uses tags to mark-up text for presentational purposes in web pages, XML is used to specify information structure. XML and its related technologies provide a powerful and very flexible way of encoding information. Because of its transparency XML is becoming ubiquitous in modern information systems, both as a means for defining, through XML Schema Definition, and containing data. Increasingly XML Schema is being used to specify metadata systems (EML) and data transfer and exchange formats, for example ABCD and GML. Even if systems that are not defined in XML or XML Schema, it is likely that they can be expressed in one or other of these forms and that if expressions are not yet available then they will be soon. For example, neither FGDC nor DCMI are defined using XML Schema, however, XML variants of both are freely available.

A number of the systems described below are heavily dependent on XML and XML Schema and it is important to understand how these differ and complement each other. Indeed XML forms the core of interoperable Web Services.

> **XML** is a meta-mark up language. It is text based, generic and can be used to define custom tags that have specific meaning to the application. It is concerned with content rather than presentation. For example, the HTML:
>
> ```
> <h1>This is a Heading</h1>
> ```
>
> presents the text "This is a Heading" in large bold text in the web browser. Alternatively the XML:
>
> ```
> <name>Joe Bloggs</name>
> ```
>
> defines the content "Joe Bloggs" in a `<name/>` element, but it says nothing about how this data should be presented.
>
> The tags, such as `<name/>`, are defined in a Document Type Definition document (DTD) or, latterly, in an XML Schema. An XML document can be validated to contain defined tags by validating its contents against its DTD or XML Schema Definition.
>
> **XML Schema Definition (XSD)** is a definition language (not a meta-language) and defines a "class" of XML documents. Here a "class" means a named collection of document elements that constitute a particular type of document, for example an XML file containing collections data (see ABCD below). A given example of an XML Schema's document class is known as an instance document, i.e. a document that contains data. A given instance of an XML Schema class can be validated against its Schema. XML Schema provides a richer way in which to define a class than does a DTD.

An example of an XML Schema is the ABCD definition for biological collections data exchange (see below). The XML Schema defines the structure of an instance, expressed in XML, which contains a given collection data record. This provides a powerful, extensible, rich and (relatively) transparent medium for both defining and containing data structures and data. XML and XML Schema are entirely generic. However, the ABCD XML Schema Definition and the EML Schema

Definition are entirely specific to biological collections data and ecological metadata, respectively.

XML and XML Schema Definition are pervasive throughout Web Services, described below.

The XML and XML Schema Definition specifications can be found on the W3C web site at http://w3d.org

3.3.3 Web Services - WS

A Web Service is a generic term that has been defined as:

"...a piece of business logic, located somewhere on the Internet, that is accessible through standard-based Internet protocols such as HTTP or SMTP". (Chappell and Tyler, 2002).

However, in the context of the component technologies described below they are very specific applications that:

"*... are modular applications that are self-describing and that can be published, located and invoked from anywhere on the Web or within any local network based on open Internet standards.*" (Cauldwell et al., 2001).

The WS service based architecture and their supporting technologies are the result of collaboration between key software companies, including IBM, Microsoft, Sun Microsystems and others, and agreement on specific standards. In particular WS is based on XML and the Simple Object Access Protocol (SOAP), described below. WS promise a standard-based distributed information network that many projects have aspired to over the past decade, though either bespoke or less widely adopted protocols such as CORBA. WS have been embraced by the major software players and are likely to have a profound and rapid impact on distributed information sharing and processing. WS enable not only people and organisations to share data but also allow applications to share data dynamically. WS frameworks are available from most of the major software vendors, including IBM, Sun Microsystems, Microsoft, Oracle, HP, etc.

Article 17 of the Convention on Biological Diversity states:

"1. The Contracting Parties shall facilitate the exchange of information, from all publicly available sources, relevant to the conservation and sustainable use of biological diversity, taking into account the special needs of developing countries.

2. Such exchange of information shall include exchange of results of technical, scientific and socio-economic research, as well as information on training and *surveying programmes, specialized knowledge, indigenous and traditional knowledge as such and in combination with the technologies referred to in Article 16, paragraph 1. It shall also, where feasible, include repatriation of information.”* (Source www.bidiv.org).

Furthermore, Article 18.3 states that the Convention should establish a Clearing House Mechanism *“to promote and facilitate technical and scientific cooperation.”*

The objectives of IABIN support the CBD and further add that the network should:

“1. Develop an Internet-based, decentralized network to provide access to scientifically credible biodiversity information currently existing in individual institutions and agencies in the Americas.

2. Provide the tools necessary to draw knowledge from that wealth of resources, which in turn will support sound decision-making concerning the conservation and sustainable use of biodiversity … as well as in other areas critical to development and poverty alleviation.”

WS has a lot to offer IABIN, and indeed the CBD. The above objective for information sharing within the CBD and IABIN is based on the premise that information sharing is a good thing and that it is therefore worth doing. However, the leap, from providing access to information, to knowledge extraction, sound decision making, alleviation of poverty, and sustainable use and conservation of biodiversity is large, and it does not necessarily follow by simply providing access to information. This is not to say that providing access to information through discovery mechanisms, such as metadata (see Deliverable 7.2) is not worthwhile. It is, but as well as helping individuals and organizations to find relevant information in a collaborative network, WS can provide information services – i.e. bring together relevant information from heterogeneous sources to provide input to a well defined piece of “business logic” generating a value added information product. This information (or Web) service might be very simple or very complex. WS therefore provide exactly the standards based framework for interoperable data management and sharing with which this report is concerned. Furthermore, through the medium of XML, it achieves this goal in a neutral manner, making no demands on the type of operating system or computing environment required.

The most important component in developing a WS system as part of the IABIN network is perhaps not in the understanding or implementation of the technology, but in understanding what services are needed to fulfill the network’s objectives.

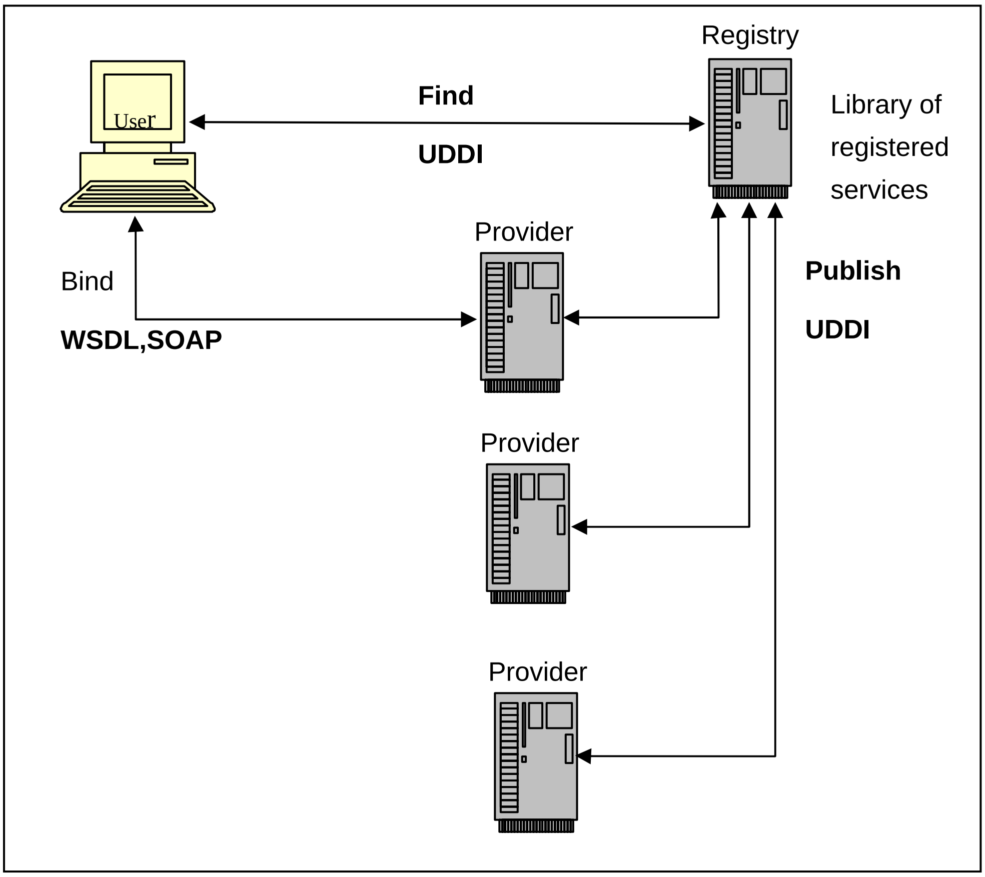


Figure 2. Service-Oriented Architecture. After Chappell and Jewel, 2002.

Figure 2. shows the service-oriented architecture of WS. To better understand WS it would be useful to consider a simple example and to examine some of the related technologies.

There is a great deal of depth in WS and space does not permit a full description of each component. The reader is referred to Cauldwell et al (2001) and Chappell and Jewell (2002). The Microsoft web site also contains useful information on WS in the form of .NET - http://www.microsoft.com/net/.

The component parts of Web Services are as follows:

1. Simple Object Access Protocol – SOAP. SOAP is a protocol for communication between computers across a network using XML. The computers need not be using the same operating systems, programming environments or database systems, and can be in a “decentralized and distributed environment”, e.g. the Internet. The protocol provides a way of encoding messages that can be transferred from one computer to another and which may contain structured data.

The acronym SOAP is derived from:

Simple – SOAP uses XML to contain data and the ubiquitous HTTP to transport it. However SOAP is necessarily complex to deal with complexity, although it does provide a transparent standard, which, if adhered to, provides interoperability;

Object – By using XML, SOAP can contain a structured data object which itself is encapsulated into a SOAP message;

Access – SOAP is independent of the transport protocol, but currently uses the ubiquitous HTTP and SMTP. By using such a common protocol SOAP can go anywhere HTTP can go;

Protocol - Using the above components SOAP comprises a protocol for the exchange of information throughout a distributed environment. (After Chappell and Jewel, 2002).

2. Web Services Description Language – WSDL. Chappel and Jewel (2002) define WSDL as:

"... an XML grammar for describing a web service as a collection of access end-points[1] *capable of exchanging messages in a procedure – or document-oriented fashion. A WSDL document is a recipe used to automate the details involved in application-to-application communication."*

Referring to Figure 1, WSDL binds the Provider with the Requester and explicitly documents the services provided by a Provider (what it does, how its functions are invoked and where it is). In order to invoke the service's functions WSDL provides an abstract description of the interface to the service.

3. Universal Description, Discovery & Integration – UDDI The last component needed to put together a Web Service is Universal Description, Discovery and Integration, or UDDI. This is an open project facilitated by www.uddi.org. It essentially provides an industry-wide directory service to enable the publication and location of Web Services. Clearly UDDI performs a metadata discover role but exposes much more detail than a simple URL and a conventional search engine will provide (Chappel and Jewell, 2002).

[1] URLs to which service requests are sent.

Figure 2 shows that the user (Requestor) can refer to the UDDI Registry to find pertinent registered services. In turn the interfaces to these services are exposed by UDDI, which in turn enables direct access to them. The Requestor can then, with the help of WSDL, understand the service's interface and via SOAP interact with it and extract data.

3.3.4 A Web Services Example

The following example illustrates the functioning of modular Biodiversity Web Service. Consider a research scientist wishing to determine the endangered plant species found at a given site, the location of specimens kept in collections, their protection status, information on the local ecosystem, and the conventions and treaties in force in the host country.

To answer these questions, access would be needed to a range of information on taxonomy, species, protected areas, ecosystems and laws. This would be a complex and possibly unachievable task without intensive research using a variety of information systems. However, breaking the problem down into modular and achievable components, the question could be answered by interrogating the following Web Services, obtained from different information providers:

- A geographic WS that provides the country name based on a coordinates;
- A conventions WS that provides the conventions and treaties signed by given countries;
- A protected areas WS that provides the protection status of a given spatial coordinate;
- A habitat WS that provides a habitat class according to the IUCN Habitats Authority file, for a given spatial coordinate;
- An endangered species WS that provides a list of endangered species for a given country, habitat and taxon;
- A collections WS that provides a list of Collections that contain samples of the given endangered plant.

Each of these individual tasks is achievable using Web Services technology and existing information which is held at a range of institutions. However the question can only be answered by interoperation with all of the institutions, using standards-based systems, i.e. WS. This is not a trivial task but is one where some components may already be available. Each of these individual services can also be put together in other configurations to answer different questions.

3.4 Protocols Specific to Biological Data

Although the tools mentioned above are generic, they can be applied effectively to biodiversity information management. This section considers some of the standards and protocols that have been developed for the specific purpose of aiding biodiversity information management and exchange. These standards are not generally incompatible with WS, although their function can overlap with WS objectives.

3.4.1 Darwin Core

The Darwin Core is part of the Species Analyst (http://speciesanalyst.net/) research project at the University of Kansas Natural History Museum and Biodiversity Research Center. As noted above, database systems that have been designed to perform similar functions, but which may be in different institutions, will differ in design. This impedes interoperability as the same SQL query applied to both systems will probably not function, let alone produce similar results. The Darwin Core aims to provide a standard common set of data elements for natural history collections, that can be mapped to bespoke systems to provide an interoperability mechanism to enable searches to be performed on heterogeneous databases.

Darwin Core therefore provides a standardised interface into natural history collections databases by using standard field names. Database table field names, data types, and domains are a type of metadata that are closely bound to the fields and data themselves, and are different to the discovery metadata discussed in report 7.2.

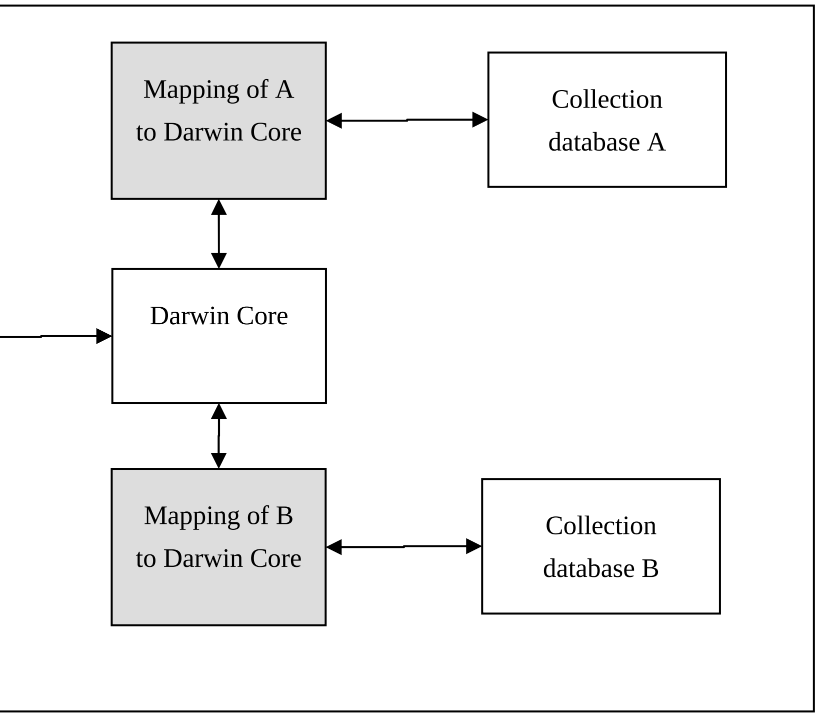


Figure 3. Darwin Core providing a common interface to heterogeneous collections databases

The common interface approach of Darwin Core has some similarities with ODBC and JDBC but operates at a different level.

The Darwin Core is concerned with natural history collections data and can be found at the Species Analyst web site at http://speciesanalyst.net

3.4.2 Access to Biological Collections Data - ABCD

ABCD is a joint working group of TDWG and the Committee on Data for Science and Technology (CODATA) and has been put forward as a TDWG standard. The objective of the working group is to:

- Foment standardization of the terminology used to model biological collection information;
- Collect and make public documents providing standards used in - or of potential use for - biological collections;
- Contribute to a general format for data exchange and retrieval for biological collections.

[Source http://bgbm3.bgbm.fu-berlin.de/TDWG/CODATA/default.htm]

In particular the ABCD working group recognises that biological collections, across the full range of biological domains, represent an immense and invaluable knowledge resource on biodiversity (2-3 billion objects), and that mobilising this under-utilised information resource is of great importance.

ABCD is concerned with describing and accessing collections databases. In this respect it is similar to the Darwin Core but at a much finer level of granularity. Indeed the relationship between ABCD and Darwin Core has similarities to the relationship between Dublin Core and EML.

ABCD is implemented as a set of XML Schemas with the following extensions:

- Botanical Gardens;
- Culture;
- Herbarium;
- Mycological;
- Palaeontological;

- PGR;
- Zoological;

The ABCD Schema can be viewed using the Schema Viewer located at: http://www.bgbm.org/scripts/ASP/TDWG/Frame.asp

3.4.3 Distributed Generic Information Retrieval - DiGIR

DiGIR is a new protocol which is designed to both replace the Z39.50 protocol, and to “retrieve structured data from multiple, heterogeneous databases”, and be easier to understand and use than Z39.50. Whilst Z39.50 is predominantly utilised in the Library domain, DiGIR has been developed specifically for the biodiversity informatics community. DiGIR is receiving an increasing amount of support and is being used by GBIF, NBII, CBIF and the ENBI. DiGIR.

3.4.4 Xanthoria

Xanthoria is a metadata query system closely associated with the Ecological Metadata Language (EML) and developed by Arizona State University. Xanthoria provides a search interface dynamically generated from the EML Schema, enabling users to either pose general queries on the whole schema or to query each element individually through a hierarchical interface.

The system uses the SOAP protocol to transport structured query requests and responses in an XML SOAP message between the query application and a set of registered metadata provider systems. The registered providers can store their metadata in conventional relational databases, native XML databases, XML files or hybrid systems such as the KNB Metacat XML database.

Xylographa, a metadata tool, is also under development.

Further information on Xanthoria can be found at the following website: http://cochise.asu.edu/bdi/Subjects/Xanthoria/ .

CHAPTER 4 EXAMPLES OF TOOLS AND BEST PRACTICE

There are many examples of information systems that aim to foster biodiversity information sharing through interoperability. The CBD CHM website lists some 28 global initiatives and 14 regional and national initiatives, although not all are directly relevant to IABIN. This chapter describes several of those that are.

4.1 Clearing House Mechanism (CHM)

The Clearing House Mechanism (CHM) of the Convention on Biological Diversity (DBD) is the most prominent biodiversity information sharing initiative. The CHM is based upon Articles 17 and 18 of the Convention, relating to the sharing of information and technical and scientific exchange. The CBD defines a CHM as an *"…agency that brings together seekers and providers of goods, services or information, thus matching demand with supply"*. More specifically the CHM's mission is to:

- Promote and facilitate technical and scientific cooperation, within and between countries;
- Develop a global mechanism for exchanging and integrating information on biodiversity;
- Develop the necessary human and technological network.

(Source: http://www.biodiv.org)

To achieve its goals the CHM has established national focal points in each country that is a signatory of the CBD, and regional and sub-regional networks are also being fostered, such as AIBIN.

The CHM provides a CHM Toolkit to help Parties to the Convention engage in the CHM process. The toolkit provides an information-sharing framework around the following headings, although it provides no specific software system:

1. Establishing CHM National Focal Points;
2. Developing a National CHM Website;
3. Technical Support and other Toolkits;
4. Partnering and funding opportunities;
5. Common Formats and Controlled Vocabularies;
6. Metadata.

The CHM has evolved as a network of HTML web sites with common goals. This was perhaps inevitable considering its evolution during the late 1990s. The CHM collaborates with other biodiversity informatics initiatives such as the Species Analyst, UNEP.net, Sistema Integrado de Información Taxonómica (SIIT) and GBIF. Some of these initiatives are well placed to assist the CHM to leverage the emerging technologies of Web Services.

The CBD CHM web site is at http://www.biodiv.org/chm/default.aspx

The CHM Toolkit is at http://www.biodiv.org/chm/toolkit/

4.2 World Biodiversity Information Network (REMIB)

The World Biodiversity Information Network, or REMIB, was established in Mexico in 1993 at Oaxaca by the National Commission for the Knowledge and Use of Biodiversity (CONABIO) and in support of the CBD. The stated purpose of REMIB is, in the first instance, to facilitate access to Mexican biological data through an information-sharing network. Interest in the network has expanded and the name has changed to the World Biodiversity Information Network, and it now includes data from some 146 countries with more than 6 million data records. These are predominantly taxonomic, but there are plans to include cultural, ecological, cartographic, bibliographic and ethno-biological data as well.

The REMIB system has a clear data access policy, which users must accept before they use the system.

Further information on REMIB can be found at:
http://www.conabio.gob.mx/remib_ingles/doctos/acerca_remib_ing.html
Further information on CONABIO can be found at: http://www.conabio.gob.mx

4.3 CGIAR System-wide Information Network for Genetic Resources (SINGER)

The CGIAR System-wide Information Network for Genetic Resources, or SINGER, is a classic information sharing and database interoperability initiative in biodiversity. In the SINGER case the “system” is the network of CGIAR research centres (including CIAT, CIMMYT, CIP, ICARDA, ICLARM, ICRAF, ICRISAT, ILRI, IITA, IPGRI/INIBAP, IRRI, WARDA), which collectively hold germplasm collections of more than 500,000 sample species. SINGER operates by linking the genetic resources information systems of each relevant CGIAR centre, enabling them to be searched collectively.

It is understood that SINGER collates the contents of the participating collections in a central database repository, with appropriate standardisation, rather than performing distributed searching across each centre’s database system.

SINGER can search by:

- Taxonomy;
- Collecting mission;
- Accession;
- Material transfer;
- Cooperating Centre;
- Characterisation and Evaluation

Further information on SINGER can be found at http://www.singer.cgiar.org

4.4 Global Biodiversity Information Facility

The Global Biodiversity Information Facility is an international non-profit organisation with a mission to:

" ... make the world's primary data on biodiversity freely and universally available via the Internet", with the benefits of contributing to *"economic growth, ecological sustainability, social outcomes and scientific research"*.

GBIF does not compete with other similar initiatives but instead seeks to work in support of activities such as the Clearing House Mechanism, the Global Taxonomic Initiative and the Convention on Biological Diversity, amongst others. It should be noted that GBIF's focus is currently predominantly on taxonomy and specimen collections, as evidenced by the search and browse features listed below.

Technologically GBIF has embraced Web Services and Grid technologies and is rapidly establishing itself as a leading initiative in the field, including a global metadata registry of available biodiversity data. Capitalising on this global register, GBIF will provide distributed search mechanisms to access these databases. GBIF envisages that in the long term it will include molecular, genetic, ecological and ecosystem level databases as part of its network (Source http://www.gbif.org).

GBIF utilises a number of the tools and protocols described above, namely the Darwin Core, ABCD, DiGIR and UDDI.

At the time of writing the GBIF Data Portal has 33 data providers which provide some 11,000,000 data records. These can be browsed/searched by:

- Browse facilities by taxonomy;
- Browse by data providers of major taxonomic databases;

- Browse by data providers that provide specimen / observation records – a simple metadata listing;
- Search by common and scientific names.

4.5 Species Analyst

The Species Analyst is a research project based at the University of Kansas, which aims to develop standards and software tools to access natural history collections databases. The Species Analyst network is based upon the Darwin Core and Z39.50 search protocol discussed elsewhere. A number of projects are closely associated with the Species Analyst project including the DiGIR protocol development which aims to replace Z39.50 as a protocol for biological data.

The SA network currently provides searches across some 120 natural history collection databases.

Further information on the Species Analyst project can be found at http://speciesanalyst.net/

4.6 Knowledge Network for Biocomplexity (KNB)

The Knowledge Network for Biocomplexity, or KNB, provides a comprehensive metadata language for Ecology, known as the Ecological Metadata Language (EML). This is accompanied by a number of tools to enable the compilation, storage and searching of metadata. Furthermore, the language provides a comprehensive description of data themselves and maintains a close binding between data and metadata.

Like many similar systems the KNB web interface (http://knb.ecoinformatics.org) provides both a browsing interface, based on taxonomy, level of organisation, ecology (including biodiversity), measurements, evolution and habitat, and a simple search interface. These interfaces allow searching of the system's metadata to locate appropriate Data Packages (metadata and data), which may include links to the actual data or web contents of the provider.

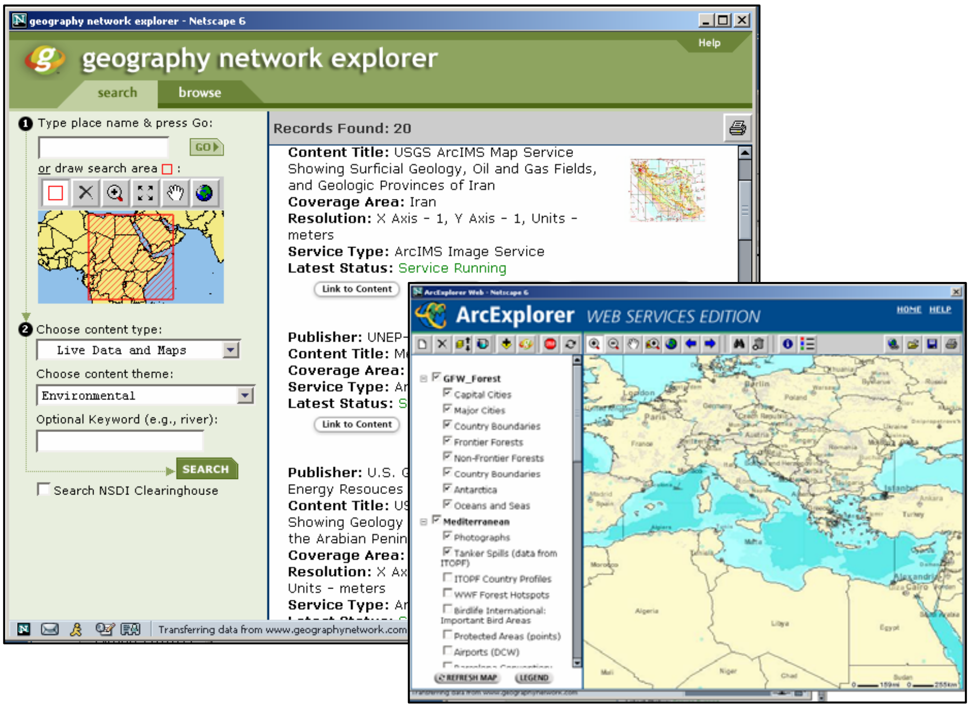
The richness of the EML language provides a detailed description of information resources that goes beyond discovery level metadata and which can provide deep searching into datasets given the appropriate tools. The suite of KNB tools, founded on EML and including Xanthoria, are still being developed but hold promise to provide complementary functionality to systems targeted at collections data, such as ABCD.

4.7 Geographynetwork.com

The Geography Network, http://www.geographynetwork.com is a commercial endeavour by ESRI, the leading Geographic Information Systems software provider. It is included here not because it is particularly relevant to biodiversity information but rather because it demonstrates how geospatial data can be provided across the Internet in an information-sharing context. The geospatial community has been at the leading edge of developing Internet standards and protocols for information-sharing. ESRI's products are based on proprietary systems but can also support OGC compliant protocols.

The Geography Network enables data providers to publish Web based GIS services, (applications, data, maps and web services) to a central metadata repository and to expose them to the user community through a standard interface that enables searching for maps by spatial extent, content type and theme.

Map data located by the service can be displayed and layers combined from different sources in a common browser-based viewer. The early implementation of the Geography Network inspired the development of UNEP.net.



Records Found: 20

Content Title: USGS ArcIMS Map Service Showing Surficial Geology, Oil and Gas Fields, and Geologic Provinces of Iran
Coverage Area: Iran
Resolution: X Axis - 1, Y Axis - 1, Units - meters
Service Type: ArcIMS Image Service
Latest Status: Service Running

Publisher: UNEP-
Content Title: M
Coverage Area:
Service Type: Ar
Latest Status: S

Publisher: U.S. G
Energy Resouces
Content Title: US
Showing Geology
the Arabian Penin
Coverage Area:
Resolution: X Ax
Units - meters
Service Type: Ar

Figure 4. Geography Network search and visualisation interface.

These functions are purely representational of what is where. However, spatial Web Services based services are also becoming available that provide specific service-based data on demand.

An increasing number of open source tools, based on the OGC standards, are becoming available to perform similar tasks.

4.8 UNEP.net

UNEP.net provides a global portal to environmental information categorised by themes and regions. The aim of the portal is to bring together environmental knowledge from the distributed UNEP office as a coherent whole. The portal provides access to thematic portals to a range of environmental themes, including climate change, freshwater, GEO, mountains, socioeconomics and the urban environment. Within these thematic categories, biodiversity features as a strong sub-theme. The system is based on an underlying metadata system that provides search facilities similar to the Geography Network. UNEP.net provides an access portal to a formidable range of data, much of it spatial in nature, but only some of it relating to biodiversity. As such it provides a useful data sharing function but its potential as a system for database interoperability is currently limited.

Further information on UNEP.net is located at http://www.unep.net#

CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This report covers a diverse, and in parts as yet somewhat esoteric, range of information systems tools for information sharing and database interoperability. The W3C's "grand plan" for the development of the WWW into a "Semantic Web", in which content semantics is more important than content presentation, is still being developed. These new WWW tools, such as XML, XML Schema Definition, SOAP, WSDL, UDDI, etc., are likely to become very common in business-to-business computing in the coming years as they have been adopted by all of the major software players, including IBM, Microsoft, Sun Microsystems and Oracle. Furthermore, these new systems are being taken up rapidly by both commercial software developers, such as ESRI, standards organisations, such as TDWG and OGC, and biodiversity information networks, such as GBIF. By embracing the Web Services' "service architecture" these initiatives are laying the foundation of an *e-business for biodiversity*. However, Web Services are still very new and their full potential has yet to be realised, and it is certain that there will be many obstacles to overcome in realising that potential. To date their utilisation in the biodiversity community has been limited, but these are early days.

Perhaps the greatest challenge is not in adopting the technologies, but in understanding how they can best be utilised in the biodiversity context. The simple WS examples sited in this report are intended to stimulate this debate rather than be a definitive biodiversity application. The full potential of WS for biodiversity can only be realised through a sound analysis of information requirements and design of services to address these requirements. As things stand the information sharing objectives of many initiatives are actually very vague.

A number, but by no means all, of the information sharing initiatives at the international level have been briefly reviewed. Some of these initiatives are clearly based on traditional information sharing and techniques; others are based on bespoke software and others still are based on standards-based Web Services.

There is clearly a tendency for information systems containing similar content to form sub-networks and to be presented together, for example germplasm collections. Whilst cross searching such collections is very useful, it does not enable complex questions to be addressed. The service model, with different information providers providing targeted and interoperable services, has the potential to deliver more that the sum of the respective parts.

5.2 Recommendations

The IABIN Portal Architecture report (McClarty, 2003) makes the following recommendations:

1. IABIN should adhere to an open technology policy that is implementation agnostic;

2. IABIN should be committed to open standards that foster interoperability;

3. Software developed in IABIN should be shared freely;

4. IABIN should work closely with the GBIF initiative.

The analysis presented in this report supports these recommendations, and suggests that the following should be added:

5. IABIN should also work closely with other initiatives that include not only taxonomic information but also information on species, protected areas, ecosystems and responses. The linkages with these initiatives should also consider utilising Web Services technologies if they do not already do so;

6. As well as utilising current and emerging information systems standards, IABIN should also consider appropriate Web Services that can be provided in addition to cross database searching. These modular (value added) applications should comprise information components within the greater biodiversity policy context;

7. To achieve 6, above, a design exercise should be initiated to develop suitable Use Cases which identify the relevant system (actors) and functional relationships between them (use cases).

CHAPTER 6 REFERENCES

Biodiversity Conservation Information System, 2000, Framework for Information Sharing, Busby, J., R. Ed.

Chappell, D., A., Jewel, T., 2002, Java Web Services, O'Reilly, Sebastopol, CA.

Caldwell, P., Chawla, R., Chopra, V., Damschen, G., Dix, C., Hong, T., Norton, F., Ogbuji, U., Olander, G., Richman, A., Saunders, K., Zaev, Z, 2001, Professional XML Web Services, Wrox Press, Birmingham, UK.

McClarty, D., 2003, IABIN Portal Architecture, IABIN GEF PDF Project Report, Version 0.5 July 2003. http://www.iabin.net.

ANNEX 1 - Key Contacts

Standards, Information Models, and Data Dictionaries
for Biological Collections

http://www.bgbm.org/TDWG/acc/Referenc.htm

ANNEX 2 - Acronyms and Abbreviations

CBD	Convention on Biological Diversity
CGIAR	Consultative Group on International Agricultural Research
CODATA	Committee on Data for Science and Technology
CONABIO	National Commission for the Knowledge and Use of Biodiversity
CORBA	Common Object Request Broker Architecture
DTD	Document Type Definition
EML	Ecological Metadata Language – an XML dialect
ER	Entity-Relationship
GIS	Geographic Information Systems
GML	Geographic Markup Language
HTTP	HyperText Transfer Protocol
JDBC	Java DataBase Connectivity
KNB	Knowledge Network for Biocomplexity
ODBC	Open DataBase Connectivity
REMIB	World Biodiversity Information Network
SINGER	CGIAR System-wide Information Network for Genetic Resources
SQL	Structured Query Language
SMTP	Simple Mail Transfer Protocol
TDWG	Taxonomic Database Working Group
UML	Unified Modelling Language
XML	eXtensible Markup Language
XSD	XML Schema Definition

ANNEX 3 - Glossary of Terms Used

Term	*Definition*
Metadata	"Data about data"